This paper presents a numerical experiment to estimate the relative influence of the different sampling periods in the estimate of the fraction of young water (fyw). The authors used 1-year long subsets of precipitation and stream tracer data sampled sequentially over a 4.5-year long record. This resulted in 189 different to estimate fyw based on sine function fits. I find the paper interesting as this approach to estimate the event water fraction is becoming popular among hydrologist. However I dough this paper provides information useful outside the catchment where the data was collected. The authors made no case on how these findings would be relevant to other locations. As such, it reads like a case study. Therefore, I suggest this paper not be consider for publication in HESS in its present form. In addition, I found the study lacks proper justification for the used of 2% difference in fyw as indicative of a significant difference.

We thank reviewer #1 for the helpful comments.

Usefulness to other catchments

The main aim of this study is to present a *generic method* to analyze the time-variance of the fraction of young water. The reviewer already mentioned that Fyw is becoming more popular. Still, we lack information on when to best use this method and how sensitive it is to different datasets (e.g. frequency of sampling and length of observation time). Thus, investigating its use, limits and pitfalls is very important before we apply it to any catchment and particularly when comparing results from different catchments. Many catchment studies showed that the transit time of water strongly varies (e.g. Harman, 2015; Heidbüchel et al., 2013), and it is thus very likely that Fyw also varies in other locations than ours.

While previous studies focused on hydroclimatic and methodological influences on Fyw, this study is the first to focus on the influence of the sampling period and length. This is a first step, and it is highly recommended that this is repeated in other catchments to assess if this is a general situation or only a few catchments have time-varying Fyw (which we doubt because of strongly varying transit times in general).

Ultimately, catchment comparison studies that rely on Fyw should be based on comparable Fyw results. For example, Stockinger et al. 2016 already showed that only changing the sampling frequency of isotopes data led to drastically changed Fyw results. The present study goes further and shows that also the sampling period can influence Fyw. Based on reviewer comments, we now also present the associated uncertainties. Information on the various influences on Fyw results is critical for catchment comparison studies. We encourage hydrologists to use our generic method to test the existence of strong time-variances of Fyw in other catchments. However, the application of this method to a large set of catchments is beyond the scope of this study.

Additionally, we now conducted a preliminary extended analysis on possible hydro-climatological influences on Fyw and its uncertainty to further foster transferability to other catchments. The (hydrologic, meteorological and isotopic) data suggests that unique meteorological conditions might have influenced the Fyw estimate.

We will expand the revised manuscript with these discussion points.

Using 2% as difference

We now applied Gauß error propagation to estimate uncertainty in Fyw and will use it to derive a data-driven threshold instead of 2% difference (preliminary Figure R1):

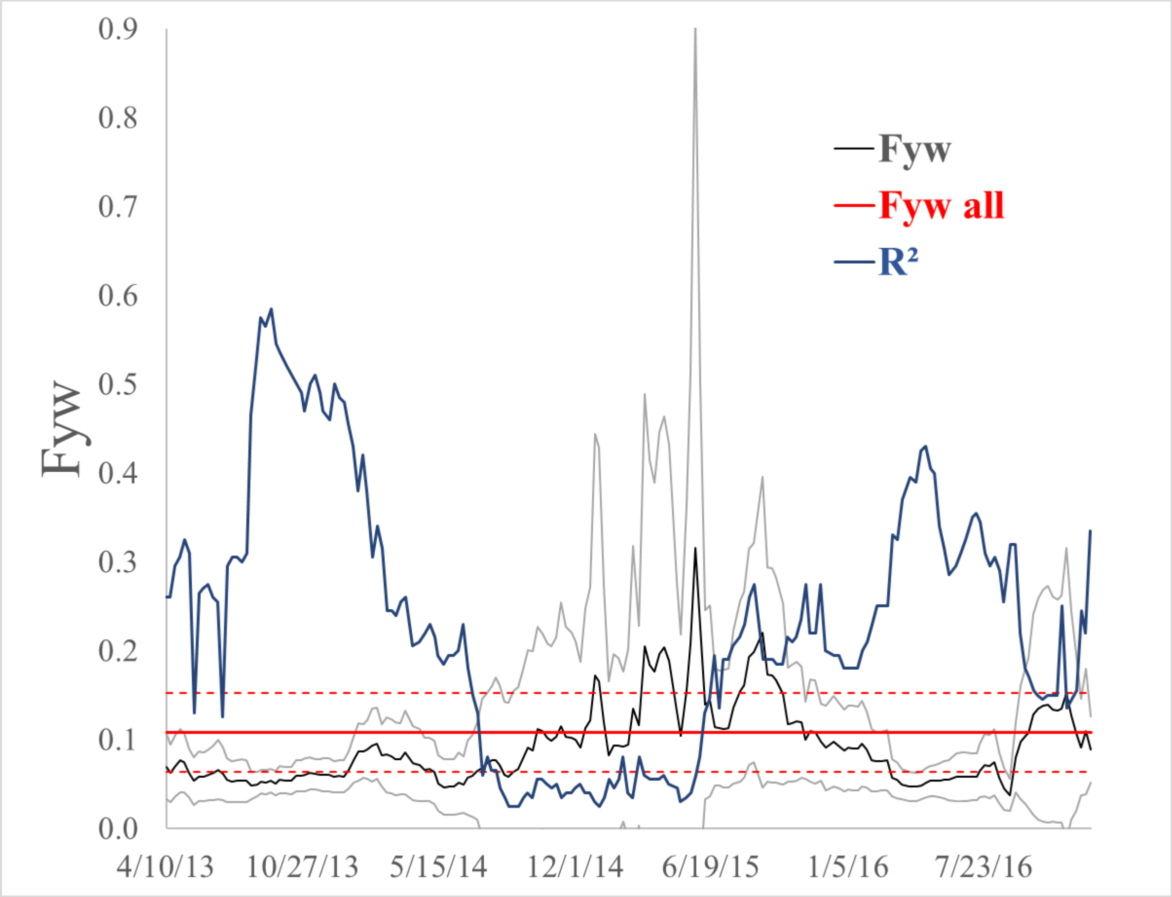

Figure R1. 189 Fyw results (black) and uncertainty (grey) compared to Fyw for all data (red, solid line) and respective uncertainty (red, dashed line). Additionally plotted is the adjusted R^2 (blue).

The following can be said from this result:

a) with a drop in R^2 below approx. 0.2 the uncertainty increases drastically. This, together with the strongly fluctuating Fyw results (page 11, lines 6-8), indicates that in the Wüstebach an R^2 of at least 0.2 should be reached. We highly recommend conducting similar studies in different catchments to test whether different R^2 threshold values exist in other catchments.
b) Fyw of all data ("Fyw all" in Figure R1) had an uncertainty of appr. ±4%. We will use this new data-driven value instead of the ±2% for re-evaluating our hypotheses.

Thus, we will use a threshold value based on the information contained in our data instead of a value chosen from studies of different catchments (i.e., ±2% taken from Lutz et al., 2018).

The statistical approach is also somehow vague. For example, it would be important to know how does the r2 fits of the input compare to the r2 fits of the output. This would allow understanding what is driving to low mean r2 values that were observed for some of the results.
A comparison of input and output R^2 fits is shown in Figure 4: the mean R^2, as well as R^2 of throughfall (input to the catchment, TF in Figure 4a) and R^2 of streamflow (output from the catchment, Q in Figure 4a). The sentence of page 8, line 4 might have been misleading. We suggest the following new sentence:
"The mean adjusted R^2 is the arithmetic mean of the precipitation and the streamflow adjusted R^2 values of the respective sine waves. It showed a marked decrease[…]"

As both R^2 for TF and Q drop significantly (Figure 4a), we assumed that the largest influence on the low R^2 values were the low amplitudes of the sine waves. A comparison of TF and Q amplitudes is shown in Figure 4b, where low TF amplitudes seem to drive low R^2 values.

In the manuscript we only compared the mean amplitude. We will add a discussion about the suggested influence of TF and Q amplitudes on low R^2.

Other specific comments:
Line 21 (P1): Sentence in poorly worded.
Based on the new uncertainty analysis, we suggest:

"Our results showed a high short-term variability and increase in uncertainty of Fyw when the mean adjusted R^2 was below 0.2. Consequently, a low R^2 indicated highly uncertain Fyw results"

Line 23 (P1): The abstract indicates that they recommend an r2 threshold for future studies. However nowhere in the text, the authors offer any justification for the limit.
The added uncertainty estimate of Fyw supports a certain threshold of quality for the fit. Very low goodness-of-fit values increase the uncertainty (as expected). Otherwise, R^2 values close to 0 would also be accepted and the respective Fyw results accepted as they are.

In a revised version we would expand on the currently existing discussion (page 11, line 11. Added text in **bold**):

"[…], we assumed that **in our case** the Fyw calculation method reached its limit below an average adjusted R^2 = 0.2. Fyw became highly sensitive to a small change in input data and highly uncertain. Additional investigations on the sensitivity of Fyw to the goodness-of-fit (not necessarily only measured with adjusted R^2) are subject to future studies. **It remains to be seen if a value of 0.2 for adjusted R^2 is a critical threshold value for Fyw or if other studies in different catchments show varying results.**"

Line 6 (P2), Line 15 (P13) and elsewhere: Better to refer "water stable isotopes" rather than "stable isotopes of water"
We will change this.

Line 16 (P3): Indicates that the hypotheses were tested against rules of acceptance that were based on whether differences in Fyw exceeded a threshold value of ± 2%. A more comprehensive justification for the 2% threshold should be included.
We will change this to the data-based 4% (see answer above)

Line 17-18 (P4): Please explain how this precision was estimated. Did you collect duplicate samples?
This is the long-term precision derived from the uncertainty of 10,000s of measurements of various water samples conducted during the last years. Each unique sample is measured 6 times.

Line 21 (P4): Did you consider using deuterium instead of 18O?
D and 18O are strongly correlated (R^2 = 0.98) so we did not consider using it.

Line14-16 (P5): It would be interesting to see the distributions or R2 of both fits independently.
R^2 is shown in Figure 4a for both throughfall and streamflow (orange lines labeled TF and Q).

Line 1-5 (P6): Since the 2% threshold is mentioned in the introduction this explanation belongs there.
We will move the suggested sentences and adapt them to the new threshold value.

Line 22 (P7): These values are very low. An r2 =-.08 would indicate that a sine wave function is weak to describe the variability of the data.
The low R^2 mentioned are values for the single sine wave fit to the full 4.5-year time series and are 0.08 for TF and 0.2 for Q. We fully agree that those sine wave functions are gross simplification of the

inter-annual variability of isotopes; which is the point of this study: a single sine wave fit oversimplifies naturally occurring, annual variations. However, even if a sine wave is weak to describe the data, the Fyw calculation method is based on using the amplitudes of sine wave functions (Kirchner, 2016). We will add this discussion in the revised version.

A quick way of assessing inter-annual variability is to use the moving average over isotope values. The following two figures show the moving average (12 week interval, not weighed, thick black line) plotted over the single sine wave fit and the 189 fits. The 189 time-variable sine waves are better suited (but not perfect) to capture the inter-annual variability than the single sine wave. We do not intend of using these figures in a revised manuscript version.

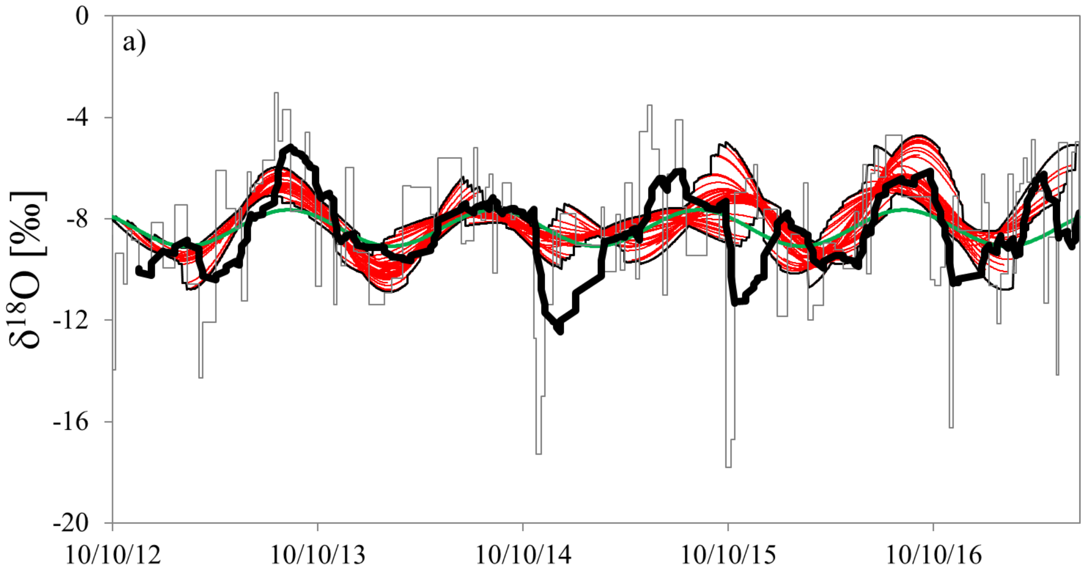

Figure R2. 12-week, non-weighed moving average over the throughfall isotope data, compared to the 189 1-year sine waves (red) and the single 4.5-years sine wave (green).

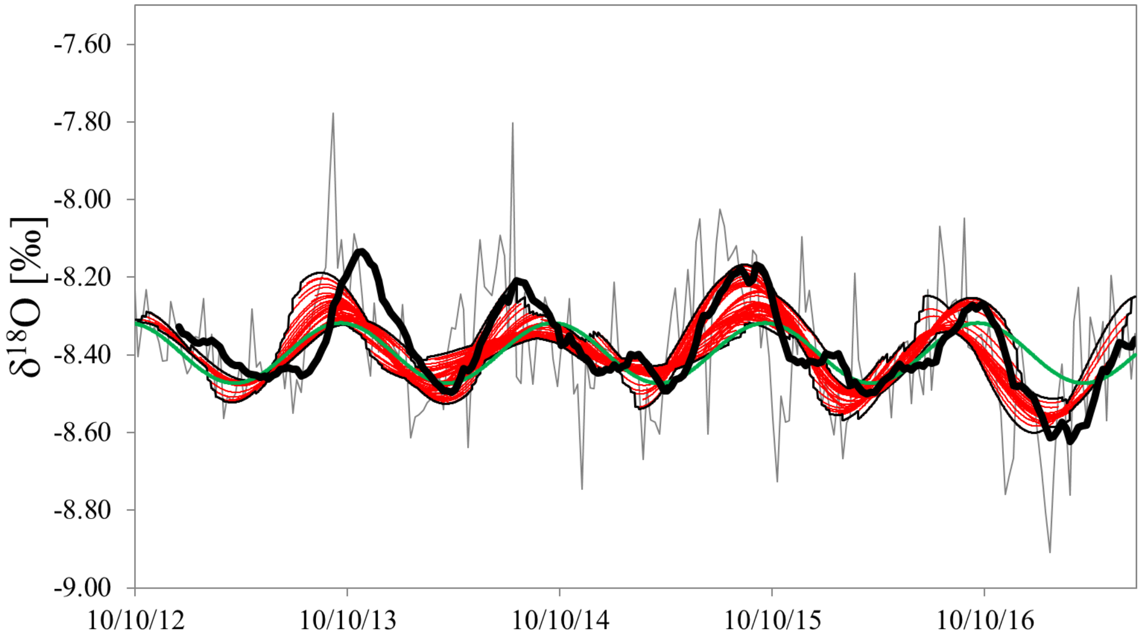

Figure R3. 12-week, non-weighed moving average over the streamflow isotope data, compared to the 189 1-year sine waves (red) and the single 4.5-years sine wave (green).

Line 24-25 (P7): Can you provide information about the range of the r2 of these fits

We will add:
Precipitation min, max and mean R^2: -0.03, 0.63, 0.21
Streamflow min, max and mean R^2: 0.00, 0.55, 0.25

Line 4-9 (P8): Would this indicate that the sine fit method is not appropriate for much of 2014? How confident can one be of the Fyw estimates when the r2 are below 0.2?
We added the uncertainty estimate (see above, Figure R1) and Fyw indeed becomes highly unreliable in this period. We encourage studies of the reliability of Fyw based on goodness-of-fit measures of the sine waves.

Line 12-14 (P8): It is not clear what is the significance of this clustering of points.
We will remove this sentence, as this was just an observation on our part.

Line 19 (P8): Considering how skew the data is would it be better to use the median? Also I suggesting some standard deviation or standard error.
The median value is approximately 8.0% Fyw (complete data set, including the low R^2 period). We will add the median as another measure in the text.

We will incorporate the error estimates obtained by Gauß error propagation.

Line 20 (P8) Please consider some measure of error or uncertainty in the fyw estimates.
Please see the answer and new Figure R1 above.

Line 23-345 (P8): Please elaborate, that is indicate how many of the 189 were between this ranges.
Line 20 mentions that 63 results from 189 were within the range. The reduced data set of Fyw (leaving out the low R^2) has a total of 124 Fyw values, with 66 being in the range (53%). We will mention those numbers in the manuscript.

Line 29 (P8): Please provide some statistical information about the strength of the correlation.
The equation for the whole data set (including the low R^2) is
Runoff Coefficient = -1.24 * Fyw + 1.13, with an adjusted R^2 of 0.30 and p-value of 3E-16

The equation for the limited data set (low R^2 excluded) is
Runoff Coefficient = -2.11 * Fyw + 1.19, with an adjusted R^2 of 0.23 and p-value of 2E-7

We will add this information to the revised text.

Line 10 (P9): This is confusing about figure 9. Are these the 189 fits? That is, are these fits over a one-year duration time series?
We agree that the sentence is confusing. These are indeed the 189 fits.
We suggest adapting the sentence "As mentioned in the methods, the Fyw results were put in the middle of the one-year calculation period (calculating from February 2016 to February 2017, the result would be displayed as a data point in August 2016). We grouped together all Fyw results that were assigned to a specific calendar month to detect possible seasonality."

Line17 (P9): Where is the value of d18O for ground water coming from?
The Wüstebach catchment is extensively monitored and groundwater d18O data is available. We did not use this data at first but will add it to the revised version.

Line 19-20 (P9): please elaborate some more in the parallel to the Weigand et al. [2017] study
We extended the sentence:

"The study by Weigand et al. [2017] came to the same conclusion for the Wüstebach catchment using wavelet analysis of nitrate and DOC data collected at mainstream and tributary locations. While lower altitude locations of the catchment near the outlet were dominated by groundwater, higher altitude areas were less affected. This finding was additionally supported by field observations of shallow groundwater."

Line 18-19 P13): How do we know this conclusion is relevant to other catchments?
We expect that changes in flow paths over time alter the Fyw result of a catchment. A one-year long time series of tracer data might lead to very different Fyw results depending on when the tracer data was sampled. Thus, the results of the comparison study could vary greatly simply by shifting the one-year sampling window by a couple of weeks. We suggest using the ensemble of Fyw values and their distribution to get an additional uncertainty estimate of Fyw. We'll add this information to the revised manuscript.

Please see also our answer to the major comment regarding usefulness of our results for other catchments.

Line 24-25 (P 13): This sentence is vague. Please explain.
We adapted the sentence to make it clearer:

"If feasible, we recommend investigating a multi-year time series of tracer data with the method suggested in this study. That is, to use a one-year moving time window and estimate an ensemble of Fyw results to derive its uncertainty."

Line 26-28 (P13) It would be important to understand if the variability observed here would be relevant to understand difference across catchments.
In a revised version we will phrase this more clearly.
As discussed above, different sampling periods yield (occasionally very) different Fyw result that would influence the interpretation of catchment comparison studies. Please also refer to the discussion points above regarding the usefulness to other catchments.

Figure 1: The markers for the precipitation and runoff gauges are too similar. Add latitude and longitude grids to the map. The contour should be in the legend indicating the units. In addition, the font next to the contours is difficult to read.
We will adapt the figure accordingly.

Figure 4: In the legend, please clarify that "mean" refers to
We will change it to "Mean R^2", "TF R^2" and "Q R^2" to clarify.

Figure 7: Please add a legend.
We will do this.

Figure 8: The caption should include an explanation of what is hypothesis 2
We will add the explanation.

References

Harman, C. J. (2015), Time-variable transit time distributions and transport: Theory and application to storage-dependent transport of chloride in a watershed, Water Resour. Res., 51, doi:10.1002/2014WR015707.

Heidbüchel, I., P. A. Troch, and S. W. Lyon (2013), Separating physical and meteorological controls of variable transit times in zero-order catchments, Water Resour. Res., 49, 7644–7657, doi:10.1002/2012WR013149.

Kirchner, J. W.: Aggregation in environmental systems - Part 1: Seasonal tracer cycles quantify young water fractions, but not mean transit times, in spatially heterogeneous catchments, Hydrol Earth Syst Sc, 20(1), 279-297, 2016.

Lutz, S. R., Krieg, R., Müller, C., Zink, M., Knöller, K., Samaniego, L., and Merz, R.: Spatial patterns of water age: Using young water fractions to improve the characterization of transit times in contrasting catchments. Water Resources Research,54, 4767–4784.https://doi.org/10.1029/2017WR022216, 2018.

Stockinger, M.P., H.R. Bogena, A. Lücke, B. Diekkrüger, T. Cornelissen and H. Vereecken (2016): Tracer sampling frequency influences estimates of young water fraction and streamwater transit time distribution. J. Hydrol. 541: 952-964, doi:10.1016/j.jhydrol.2016.08.007.

Weigand, S., R. Bol, B. Reichert, A. Graf, I. Wiekenkamp, M. Stockinger, A. Lücke, W. Tappe, H. Bogena, T. Pütz, W. Amelung and H. Vereecken (2017): Spatiotemporal dependency of dissolved organic carbon to nitrate in stream- and groundwater of a humid forested catchment – a wavelet transform coherence analysis. Vadose Zone J. 16(3), doi:10.2136/vzj2016.09.0077.